The Korean Journal of Applied Statistics
2024, Vol. 37, No. 5, 541–551

DOI: https://doi.org/10.5351/KJAS.2024.37.5.541
Print ISSN1225-066X / Online ISSN 2383-5818

A statistical journey to DNN, the first trip:

From regression to deep neural network

Hee Ju Kim[a], In Jun Hwang[a], Yu Jin Kim[a], Yoon Dong Lee[1,a]

[a]Business School, Sogang University

Abstract

It has become difficult to discuss statistics without mentioning recent advancements in artificial intelligence and deep neural networks. While the progress in artificial intelligence and deep neural networks is also a result of major research achievements in statistics, modern statistics and artificial intelligence are often perceived as distinctly different approaches. The primary reason for this seems to be that the statistics education curriculum has not evolved to keep pace with the times. In this paper, to establish a framework for the expansion and development of statistics education, we examine the relationship between deep neural networks, specifically multi-layer perceptrons, and regression analysis from a statistical perspective, and explore their similarities and differences.

Keywords: deep neural net, classification, multi-layer perceptron, regression, generalized linear model

1. 서론

인공지능 기술을 언급하지 않고 현대통계학을 거론하기 어려운 시대가 되었다. 본 논문에서는, 인공지능 기술, 그 중에서도 심층신경망(deep neural network)을 통계학의 관점에서 살펴보고, 심층신경망의 특징과 전통적인 통계적 방법의 차이를 비교하여 살펴본다. 본 논문은 책임저자가 여러 관련 학회 등에서 2019년 이후 구두로 발표했던 내용을 문헌으로 정리한 여러 편 중 첫 편이다.

인공지능은 다양한 분야로 응용되고 있고, 현실에서 큰 영향을 미치고 있다. 인공지능의 발전에는 컴퓨팅 이론, 의사결정 이론, 최적화 이론, 수치해석 등등의 중요 이론들이 밑받침이 되고 있다. 그 중에서도 통계학은 인공지능의 발전을 이끌어온 중추적 역할을 담당하고 있다. 인공지능의 기계학습 이론은 통계학의 한 분야로 발전해 왔고, 심층신경망(DNN)은 발전된 형태의 통계모형 그 자체다 (Hastie 등, 2001; Venables와 Ripley, 2002).

본 논문에서는, 심층신경망에서의 대표적인 모형인 다층퍼셉트론(multi-layer perceptron)을 회귀모형과 비교하여 그 공통점과 차이점을 살펴보는 방법으로, 인공지능에서의 접근법과 통계학에서의 접근법을 비교하여 살펴본다. 이로부터, 통계학 교육의 확장과 전환에 필요한 기틀을 잡고, 다양한 응용 연구를 진작할 수 있는 기회를 마련하고자 한다.

최근의 인공지능 발전을 이끈 기술적 핵심 요소는, 확률적 경사하강법(stochastic gradient descent method)으로 대표되는 수치적 함수최적화기법의 발달, 그래픽처리장치(graphic processing unit)의 발달로 대표되는 하드웨어 계산 성능의 발달, 만능변환(universal transformation)으로 대표되는 수리적 모형화 도구의 발달을

[1]Corresponding author: Business School, Sogang University, PA 804, BaekBumRo, Mapo, Seoul 04107, Korea. E-mail: widylee@sogang.ac.kr

Published 31 October 2024 / journal homepage: http://kjas.or.kr

꼽을 수 있다. 본 논문에서는 만능변환을 통한 회귀모형의 일반화라는 관점에서, 회귀분석의 발전된 형태로서 다층퍼셉트론(MLP)을 해석한다. 또 동일한 틀에서 심층신경망에서의 접근법과 전통적인 통계학에서 접근법의 차이에 대하여도 살펴본다.

2절에서는, 통합적 틀에서 심층신경망과 회귀분석을 표현하는 방법에 대하여 살펴보고, 3절에서는 회귀분석 방법과 신경망방법이 취하는 접근법의 차이점을 살펴본다. 이후 결론에서 통계학과 인공지능의 통섭적 발전을 위하여 필요한 추가적인 작업 방향을 정리하였다.

2. 회귀분석과 다층퍼셉트론의 통합적 표현

응용통계학에서의 가장 고전적인 주제는 회귀와 분류이다. 회귀분석은 관심변수(종속변수)의 변화를 원인이 되는 원인변수(독립변수)의 변화로 해석하는 가장 기본적인 통계모형이다. 관심변수의 생성에 개입되는 분포로 정규분포를 상정한다. 여러 변수들로 구성된 벡터변수를 관심변수로 상정하는 경우, 회귀모형은 다변량 회귀로 확장된다. 일반화선형모형(generalized linear model)은 관심변수의 생성에 개입되는 분포를 정규분포 이외의 분포로 확장하는 길을 제시한다. 일반화선형모형(GLM)을 통하여, 관심변수가 0과 1 두 가지 값만을 갖는 베르누이분포의 경우도 선형모형(회귀분석)과 동일한 틀에서 해석할 수 있는 방법이 제시된다. 로지스틱회귀모형은 베르누이분포에 일반화선형모형을 적용하여 얻어진다. 로지스틱회귀는 이진분류(binary classification)를 위한 표준방법이고, 분류와 회귀를 연결하는 교차점이다. 마찬가지로 로지스틱회귀에서 적용된 방법을 다항분포의 경우로 확장하여 다그룹분류(multiclass classification) 방법이 도출된다.

2.1. 표현과 기호

본 논문에서 사용할 수학적 기호로 다음과 같은 표현법을 사용한다. 관찰된 자료 D는 상정되는 모형에 따라 그 형태가 다를 수 있으나, 회귀분석을 대상으로 하는 경우, 관심이 되는 종속변수 y와 설명변수 $\boldsymbol{x} = (x_1, \ldots, x_p)$에 대하여, 각 사례 $i = 1, 2, \ldots, n$들에서 기록된 값들의 집합의 형태로 다음과 같이 표현된다.

$$D = \{(y_i, \boldsymbol{x}_i) \mid i = 1, 2, \ldots, n\}$$

즉, 관측 단위별로 표현된 회귀모형은, 평균이 0인 정규분포를 따르는 관측오차를 ϵ_i라 할 때, $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ 이 되고, 관측자료 전체에 대하여 표현된 회귀 모형은 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ 와 같은 형태로 표현한다. 여기서 $\boldsymbol{x}_i$는 행벡터이고, $\mathbf{y}$와 $\boldsymbol{\beta}$은 열벡터이다. 설명변수 x_j에 대한 반복 관측값 x_{ij}, $i = 1, 2, \ldots, n$을 묶은 (열)벡터는, 글자체를 달리하여 $\mathbf{x}_j$로 표현한다. 여러 개의 종속변수를 갖는 종속변수 벡터 $\boldsymbol{y} = (y_1, \ldots, y_m)$ 를 대상으로 하여, 다변량회귀분석 모형을 관측단위별 식으로 표현하면, 회귀계수 행렬을 B라 하고 평균이 0인 다변량 정규분포를 따르는 관측오차 벡터를 $\boldsymbol{\epsilon}_i$이라 할 때, $\boldsymbol{y}_i = \boldsymbol{x}_i B + \boldsymbol{\epsilon}_i$ 가 된다. 이를 관측자료 전체에 대하여 나타내면, $Y = XB + E$ 로 표현된다. 여기서 $\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\epsilon}_i$는 행벡터이고, 행렬 Y와 E의 크기는 $n \times m$, 행렬 X의 크기는 $n \times p$, 행렬 B의 크기는 $p \times m$이다.

심층신경망의 핵심 요소 중 하나는 활성함수(activation function)다. 활성함수에는 다음과 같이 로지스틱(logistic)함수, ReLU (rectified linear unit) 함수, 소프트맥스(softmax) 함수 등이 주로 사용된다. 이들 함수를 나타낼 때, 굳이 $f()$와 같은 형태로 함수의 이름이 f임을 표시할 필요가 없으므로, 대괄호 기호 []만으로 각 함수를 나타내기로 한다. 특별히 시그모이드임을 나타내야 할 필요가 있을 때는 $[]$, ReLU 함수임을 나타내야 할 필요가 있을 때는 $[]_+$, softmax 함수임을 나타내야 할 필요가 있을 때는 $[]_*$라고 나타내기로 하자. 예를 들어, 입력변수 $\boldsymbol{x}$에 대하여 반복적으로 활성함수를 적용하고 마지막에 softmax 함수를 적용하는 경우 $[[[\boldsymbol{x}]]]_*$와

같이 표현한다. 자주 사용되는 활성함수의 구체적인 형태는 다음과 같다.

$$\begin{aligned} &\text{로지스틱 함수:} && [x]_{\cdot} = e^x/(1+e^x) \\ &\text{ReLU 함수:} && [x]_{+} = \max(0,x) \\ &\text{소프트맥스 함수:} && [\boldsymbol{x}]_{*} = e^{-s}\left(e^{x_1},\ldots,e^{x_p}\right), \quad s = \log\left(\textstyle\sum_i e^{x_i}\right). \end{aligned}$$

일반화선형모형에서 관심변수의 평균 μ와 선형모형식 $\eta = \boldsymbol{x\beta}$ 사이의 관계를 연결함수 $g()$를 이용하여 $g(\mu) = \boldsymbol{x\beta}$ 와 같은 형태로 표현한다. 각 분포에 가장 잘 어울리는 연결함수가 정준연결함수(canonical link function) 이다. 평균이 λ인 포아송분포의 정준연결함수는 로그함수, 즉 $\eta = \log(\lambda)$ 이다. 이항분포(혹은 베르누이분포)의 경우는 로짓(logit) 함수가 정준연결함수다. 이항분포 $\text{Bin}(u,p)$에서 오즈(odds)는 $p/(1-p)$를 말하고, 로짓함수는 오즈(odds)에 대한 로그함수이고, 다음과 같다.

$$\eta = \log\left(\frac{p}{1-p}\right).$$

로지스틱 함수는, 이항분포에 대한 일반화선형모형에 적용되는 연결함수인 로짓(logit) 함수의 역함수이다. 심층신경망에서 $\sigma(x)$ 라는 기호로 자주 등장하는 시그모이드(sigmoid)는 로직스틱 함수의 다른 이름이다. 하이퍼탄젠트(tanh)는 시그모이드의 다른 형태이고, $\tanh(x) = 2\cdot\sigma(2x) - 1$ 인 관계가 있다. 심층신경망에서 활성함수로 하이퍼탄젠트를 사용하는 경우와 시그모이드를 사용하는 경우는, 가중치가 달라지는 효과가 발생하는 것 이외에 차이는 없다. 이 외에도 소프트플러스(softplus), Leaky ReLU, ELU (exponential linear unit), GELU (Gaussian linear unit) 등의 함수가 활성함수로 사용된다.

만능변환(universal transformation)은, 선형변환과 활성함수를 반복하여 적용하는 변환이다. 선형변환과 활성함수를 반복하여 적용하는 횟수를 깊이(depth)라 한다. 입력변수 $\boldsymbol{x}$ 에, 깊이가 3인 만능변환을 적용하여 출력변수 $\boldsymbol{y}$ 를 얻는 경우의 예를 식으로 표현하면 다음과 같다.

$$\boldsymbol{y} = [[[\boldsymbol{x}A]\,B]\,C]_{*}\,,$$

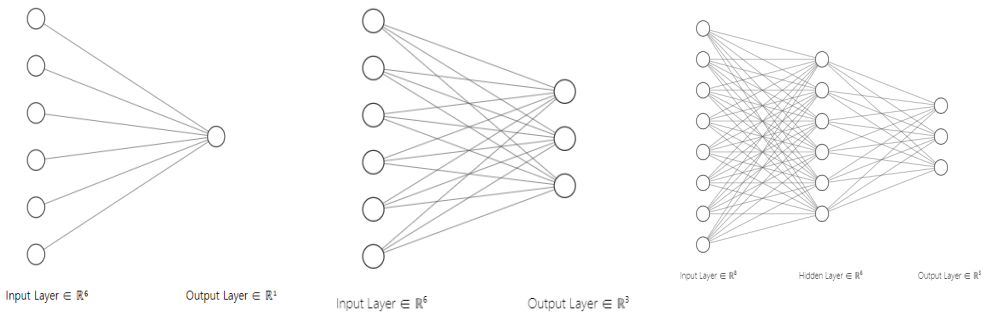
여기서 선형변환을 의미하는 가중행렬 A, B, C 는 필요에 따라 적절한 차원을 갖는 것으로 상정된다. 만능근사정리(universal approximation theorem)는 임의의 함수 $f:\mathbb{R}^p \to \mathbb{R}^m$ 가 깊이가 충분히 깊고, 각 가중행렬의 열의 개수가 충분히 크게 설정된 만능변환을 통하여 잘 근사될 수 있음을 말하는 정리이다 (Cybenko, 1989). 즉, 만능변환을 통하여, 모수적 모형화의 범주에서, 선형변환의 틀을 뛰어넘는 다양한 함수의 근사가 가능해지게 된다.

2.2. 다층퍼셉트론과 통계모형

회귀분석에서 관측값은 (다변량)정규분포를 따르는 것으로 가정된다. 즉, 설명변수를 $\boldsymbol{x}$ 라 할때, (단변량) 회귀분석에서의 종속변수 y와 다변량 회귀분석에서의 종속변수 $\boldsymbol{y}$ 는 각각 정규분포 $N(\mu,\sigma^2)$와 다변량 정규분포 $N_m(\boldsymbol{\mu},\Sigma)$를 통하여 다음과 같이 모형화한다.

$$\begin{aligned} &\text{단변량 회귀:} && y \sim N\left(\boldsymbol{x\beta},\sigma^2\right) \\ &\text{다변량 회귀:} && \boldsymbol{y} \sim N_m\left(\boldsymbol{x}B,\Sigma\right). \end{aligned}$$

이항분포 로지스틱회귀는 시행횟수가 1인 이항분포 $\text{Bin}(1,p)$에서 성공확률 p를 선형함수로 모형화하고 로지스틱함수를 활성함수로 하여 얻어진다. 마찬가지로 시행횟수가 1인 다항분포 $\text{Mnom}(1,\boldsymbol{p})$에서 성공확률



(a) Regression (b) Mnom LR (c) MLP

Figure 1: *Relation among regression, Mnom LR, and MLP.*

$\boldsymbol{p}$를 선형함수로 모형화하고 softmax 함수를 활성함수로 적용하는 것이 다항분포 로지스틱회귀이다. 이항분포 로지스틱회귀와 다항분포 로지스틱회귀(multinomial logistic regression)의 관계는 단변량 회귀와 다변량 회귀의 관계와 동일하고, 다음과 같이 표현된다.

$$\begin{aligned} &\text{이항분포 로지스틱회귀:} \quad && y \sim \text{Bin}\,(1, [\boldsymbol{x\beta}]) \\ &\text{다항분포 로지스틱회귀:} \quad && \boldsymbol{y} \sim \text{Mnom}\,(1, [\boldsymbol{xB}]_*)\,, \end{aligned}$$

이때 깊이가 3인 다층퍼셉트론은 다음과 같이 표현된다.

$$\boldsymbol{y} \sim \text{Mnom}\,(1, [[[\boldsymbol{x}A]\,B]\,C]_*)$$

이를 과정을 분리하여 다시 표현하면 다음과 같다.

$$\boldsymbol{y} \sim \text{Mnom}\,(1, [\boldsymbol{h}C]_*) \text{이고,} \quad \boldsymbol{h} = [[\boldsymbol{x}A]\,B]\,.$$

말하자면 다층퍼셉트론은 만능변환을 통하여 입력변수 $\boldsymbol{x}$를 가공하여 설명변수 $\boldsymbol{h} = [[\boldsymbol{x}A]\,B]$를 구하고, 설명변수 $\boldsymbol{h}$를 이용하여 다항분포 로지스틱회귀(Mnom LR)를 적용하는 방법이다. 전통적인 통계적 방법에서 입력변수 $\boldsymbol{x}$를 가공하는 방법은 자료분석자의 경험과 안목에 의존하여 각 자료의 특성에 따라 적당한 방법으로 가공한다. 이에 비하여 다층퍼셉트론에서는 입력변수 $\boldsymbol{x}$를 가공하는 과정을 만능변환의 형태로 모형화하고, 이 과정에 개입되는 가중치 행렬 A와 B를 적절하게 선택하는 방법을 사용한다. 심층신경망은 회귀모형이 복잡해지고 고도화된 모형이다. 신경망에서 각 노드들은 변수를 의미하고, 노드들을 연결하는 선으로 표현되는 가중치(weights) 행렬은 회귀분석에서의 회귀계수에 해당한다. 회귀분석에서 독립변수는 신경망에서의 입력변수에 해당하고, 회귀분석에서의 종속변수는 신경망에서의 출력변수에 해당한다. 회귀모형에서의 y-절편은 신경망에서의 바이어스(bias)에 해당한다. 이와 같은 사항을 간단히 표현하여 회귀분석, 다항분포 로지스틱회귀, 다층퍼셉트론의 관계를 그림으로 나타내면 Figure 1과 같다.

2.3. 손실함수

일반화선형모형(GLM)의 관점에서, (단변량) 회귀분석은 정규분포의 평균 μ에 대한 해석이고, 분류는 이항분포 혹은 다항분포의 성공확률 $\boldsymbol{p}$에 대한 해석이다. 이때 입력변수의 각 관측값 $\boldsymbol{x}_i$를, 만능변환을 통하여, 정규분포의 경우 $\mu_i = [[[\boldsymbol{x}_i A]B]C]$와 같은 형태로 평균을 해석하면 회귀목적의 심층신경망이고, 이항분포

혹은 다항분포의 경우 $\boldsymbol{p}_i = [[[\boldsymbol{x}_i A]\, B]\, C]$ 와 같은 형태로 성공확률을 해석하면 분류목적의 심층신경망이다. 분류목적의 심층신경망을 다층퍼셉트론이라 한다.

적절한 가중치 $\boldsymbol{\theta} = (A, B, C)$ 를 찾기 위한 방법으로 최대우도추정법이 고려될 수 있다. 최대우도 추정법은 음로그우도함수 $(-2)\log f(D;\boldsymbol{\theta})$ 를 최소화하는 모수 $\boldsymbol{\theta}$를 찾는 방법과 동일하다. 음로그우도함수에서 $\boldsymbol{\theta}$의 함수인 부분만을 목적함수로 설정하여 이를 손실함수라 부르고 손실함수를 최소화하는 방법이 이용된다. 손실함수를 최소화하는 과정을 '학습'이라 한다. '학습'은 '추정'을 일반화한 용어다. 정규분포를 가정하는 경우와 다항분포를 가정하는 경우에서 음로그우도로부터 얻어지는 손실함수 $L(D, \boldsymbol{\theta})$ 는 각각 다음과 같다.

$$\text{정규분포:} \qquad L(D,\boldsymbol{\theta}) = \sum_i^n (y_i - \mu_i)^2$$

$$\text{다항분포} \qquad L(D,\boldsymbol{\theta}) = -2\sum_i^n \sum_k^m y_{ik} \log p_{ik},$$

여기서 다항분포의 경우 $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})$, $y_{ik} \in \{0, 1\}$, $\sum_k^m y_{ik} = 1$ 이고, $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{im})$ 이다. 정규분포의 경우 손실함수는 '잔차제곱합'이고, 다항분포의 경우는 '크로스엔트로피'(cross entropy)이다. 크로스엔트로피 계산과정에 $\log p_{ik}$ 계산이 포함되어 있다. 최적화 과정에서 log-함수를 처리할 때 수치적 오류가 발생할 가능성이 높기 때문에 $\log p_{ik} \approx (p_{ik} - 1)$ 로 근사한 지니(Gini) 인덱스가 사용되기도 한다 (Breiman 등, 1984).

3. 회귀분석과 신경망의 차이

보통 신경망 혹은 심층신경망은 마치 회귀분석과 구별되는 방법인 것처럼 언급된다. 그 대표적인 이유들을 살펴보면 다음과 같다.

① 신경망의 블랙박스 성질
② 최적화 방법의 차이
③ 다그룹분류 포함 여부의 차이
④ 명목변수에 대한 가변수 변환 방법의 차이

그러나 위의 각 사항들을 구체적으로 살펴보면 신경망과 회귀분석의 본질적 차이라고 하기보다는 단지 적용환경의 차이에서 오는 스타일의 차이라 할 수 있고, 오히려 역설적으로 신경망과 회귀분석이 동질적 방법임을 보여주는 사항들이라 할 수 있다. 위 사항들 각각에 대하여 살펴보면 다음과 같다.

3.1. 신경망의 블랙박스 성질에 대하여

신경망 혹은 심층신경망이 회귀분석과 구별되는 다른 방법인 것처럼 인식되는 데 있어서 중요한 역할을 하는 대표적인 특징은 (심층)신경망의 블랙박스 성질이다. 그러나 신경망이 갖는 블랙박스로서의 성질은 (심층)신경망의 고유한 성질이 아니고, 회귀분석도 동일한 성질을 갖는다. 신경망에서 발생하는 블랙박스로서의 성질은 회귀분석을 통하여 보다 명확하고 쉽게 설명될 수 있다.

신경망의 블랙박스로서의 성질은, 입력변수 $\boldsymbol{x}$를 이용하여 출력변수 y를 예측하기 위하여 신경망을 사용할 때, 신경망에서 각 입력변수 x_j가 예측값 $\hat{\boldsymbol{y}}$를 생성하는 과정에서 작용하는 '가중치'(weights)에 의미를 부여하여 해석할 수는 없다고 하더라도, 신경망의 결과로 얻어지는 예측값 $\hat{\boldsymbol{y}}$은 의미있는 값으로 사용될 수 있다는 성질을 말한다.

회귀분석도 신경망과 동일한 성질을 갖는다. 회귀분석에서 회귀계수벡터 $\boldsymbol{\beta}$에 대한 최소제곱추정법에 의한 추정량 $\hat{\boldsymbol{\beta}}$은 정규방정식 $(X'X)\hat{\boldsymbol{\beta}} = X'\mathbf{y}$ 을 만족하는 근이다. 이 근은 일반화역행렬 $(X'X)^-$ 을 통하여 $\hat{\boldsymbol{\beta}} = (X'X)^- X'\mathbf{y}$ 인 형태로 표현된다. 만약 $X'X$ 의 역행렬이 존재한다면, 일반화역행렬 $(X'X)^-$ 은 역행렬

$(X'X)^{-1}$ 과 동일하고 유일하게 정의된다. 또 회귀계수 $\hat{\boldsymbol{\beta}} = (X'X)^{-}X'\mathbf{y}$ 와 예측값 $\hat{\mathbf{y}} = X(X'X)^{-}X'\mathbf{y}$ 는 유일하게 특정이 된다. 이 경우 회귀계수 각각은, 다른 설명변수의 값은 고정되어 있고 대응하는 설명변수의 값이 1 단위 증가되었을 때 예측값이 증가는 양을 의미한다.

반면 자료행렬 X 의 각 열들이 선형독립인 조건을 만족하지 않는 경우, $X'X$ 의 역행렬은 존재하지 않고, 일반화역행렬 $(X'X)^{-}$ 와 회귀계수 $\hat{\boldsymbol{\beta}} = (X'X)^{-}X'\mathbf{y}$ 는 다양한 값을 갖게 된다. 이때 얻어지는 회귀계수를 이용하여 각 설명변수의 증가가 예측값 증가에 기여하는 효과를 말할 수는 없다. 이 경우에도, $X(X'X)^{-}X'$ 은 항상 유일한 값을 갖게 되고, 예측값 $\hat{\mathbf{y}}$ 은 항상 유일하게 특정된다. 간단히 정리하여 말하자면, 자료행렬 X 의 열들이 선형독립이 아닌 경우, 추정된 회귀계수 $\hat{\boldsymbol{\beta}}$ 의 값은 고유한 값으로서 의미를 부여하여 사용할 수는 없으나, 회귀계수의 경우와 달리 예측값 $\hat{\mathbf{y}}$ 은 항상 의미있는 특정한 값을 갖게된다. 즉, 회귀분석도 신경망과 마찬가지로 블랙박스로서의 성질을 갖는다.

이는 다음과 같은 매우 간단한 예로 더욱 쉽고 명백하게 설명이 가능하다. 아버지의 키(x_1)와 성인이 된 자식의 키(y) 사이에, 다음과 같은 두 개의 회귀모형 M_1과 M_2를 고려하자.

$$\begin{aligned} M_1 : &\quad y = \beta_0 + \beta x_1 + \epsilon, \\ M_2 : &\quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \epsilon. \end{aligned}$$

회귀모형 M_1 은 단순선형회귀로 매우 일반적인 회귀분석 모형으로, 그 회귀계수 추정값 $\hat{\beta}$ 은 아버지의 키가 자식의 키에 미치는 영향으로 해석할 수 있다. 반면, 회귀모형 M_2 에는 변수 x_1 이 중복 사용된 경우로 다중공선성(multi-collinearity)의 문제가 발생한다. 이 경우 최소제곱 추정법에서 도출되는 정규방정식은 무수히 많은 근을 갖게 되어, 회귀계수 β_1 과 β_2 에 대한 추정값 $(\hat{\beta}_1, \hat{\beta}_2)$ 을 하나의 값으로 특정할 수 없다. 이 경우 $\hat{\beta}_1 + \hat{\beta}_2 = \hat{\beta}$ 인 조건을 만족하는 모든 경우가 최소제곱 추정량이 된다. 회귀계수 β_1 과 β_2 의 값으로 무수히 많은 $(\hat{\beta}_1, \hat{\beta}_2)$ 의 값 중 하나를 선택하여 거기에 특별한 의미를 부여하기는 어렵다. 그러나 이 경우에도, 회귀모형 M_2 에서 얻어지는 예측값 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1$ 은, 회귀모형 M_1 에서 얻어지는 예측값 $\hat{y} = \hat{\beta}_0 + \hat{\beta} x_1$ 과 동일하고 의미 있는 예측값으로 활용할 수 있다.

신경망이 갖는 블랙박스로서의 성질은 회귀분석과 구별되는 신경망 고유의 성질이 아니고, 입력변수들의 선택 방식에 대한 기본 가정과 입력변수 선택 방식의 차이를 말하는 것일 뿐이다. 회귀분석의 경우 자료행렬 X 의 열들이 선형독립임을 가정하는 것이 일반적인 접근법인데 비하여, 신경망의 경우 입력변수들 사이의 독립성이 깨질 수 있는 상황을 기본적으로 가정한다는 차이가 있을 뿐이다. 회귀분석에서는 회귀계수에 대한 추정량 $\hat{\boldsymbol{\beta}}$ 과 관측자료에 대한 예측값 $\hat{\mathbf{y}}$ 을 얻을 수 있는 데 반하여, 신경망은 의미있게 사용할 수 있는 회귀계수의 추정량 $\hat{\boldsymbol{\beta}}$ 을 얻는 것을 포기한다. 대신 입력변수 $\boldsymbol{x} = (x_1, \ldots, x_p)$ 를 자유롭게 선정하려는 접근법을 택한 것이다.

입력변수 사이의 독립성 가정을 포기하는 경우, 자유로운 변수 선택이 가능해지는 장점이 있지만, 계산량이 증가한다. 동일한 변수를 반복적으로 중복 선택하는 것도 허용될 것이므로, 이 경우 회귀계수 벡터 $\boldsymbol{\beta}$ 의 차원이 극단적으로 커질 수 있고 그에 따른 계산상의 부담이 초래될 수 있다. 따라서 이러한 계산상의 부담을 극복할 수 있는 효율적인 계산 방법이 담보되어야 한다.

3.2. 최적화 방법의 차이에 대하여

모수를 추정하거나 모형을 학습하기 위하여는 손실함수 $L(\boldsymbol{\theta}) = L(D, \boldsymbol{\theta})$ 를 최소화하는 과정이 필요하다. 회귀분석과 분류에서는 손실함수인 잔차제곱합과 크로스 엔트로피를 최소화하는 모수를 구한다. 손실함수를 최소화하는 최적해를 구하기 위한 기본적인 방법은 손실함수의 그래디언트(gradient) 벡터가 $\mathbf{0}$ 이 되는 해 $\hat{\boldsymbol{\theta}}$ 를 찾는 것이다. 즉, $\dot{L}(\boldsymbol{\theta}) = (\partial L(\boldsymbol{\theta}))/(\partial \boldsymbol{\theta})$ 일 때, $\dot{L}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ 인 해 $\hat{\boldsymbol{\theta}}$ 를 구해야 한다.

방정식 $\dot{L}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ 을 푸는 전통적인 방법은 뉴턴-랩슨(Newton-Rahpson) 방법이다. 뉴턴-랩슨 방법은 다음

과 같은 절차를 반복하여 손실함수를 최소화하는 최적해를 구한다.

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \left[\ddot{L}\left(\boldsymbol{\theta}^{(r)}\right)\right]^{-1}\left[\dot{L}\left(\boldsymbol{\theta}^{(r)}\right)\right], \quad r = 1, 2, 3, \ldots$$

뉴턴-랩슨 방법을 사용하기 위해서는, 해시안(Hessian) 행렬 $\ddot{L}(\boldsymbol{\theta}) = (\partial \dot{L}(\boldsymbol{\theta}))/(\partial \boldsymbol{\theta})$ 을 필요로 한다. 회귀분석의 경우, 그래디언트 벡터는 $\dot{L}(\boldsymbol{\beta}) = -2X'(\mathbf{y} - X\boldsymbol{\beta})$ 이고, 해시안 행렬은 $\ddot{L}(\boldsymbol{\beta}) = 2X'X$ 이다. 이때 $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ 이고 $X'X$ 가 역행렬을 갖는다면, $\boldsymbol{\beta}^{(1)} = (X'X)^{-1}X'\mathbf{y}$ 가 된다. 회귀분석에서 뉴턴-랩슨 방법을 적용하면 단 한 번의 반복만으로 최적해를 얻게 된다. 일반화선형모형에서는, Fisher scoring 방법을 적용하여 해시안 행렬을 그래디언트 벡터의 곱으로 근사하는 가우스-뉴턴(Gauss-Newton) 방법이 이용되고, 해시안 행렬이 역행렬을 갖는 것을 보장하기 위한 방법으로 수정된 가우스-뉴턴(modified Gauss-Newton) 방법이 이용되기도 한다.

심층신경망에서, 자료행렬 X 는 매우 큰 경우를 상정한다. 자료행렬 X 의 크기를 $n \times p$ 라 하자. 모수 $\boldsymbol{\theta}$ 의 차원 p 가 매우 커지고, 자료의 개수 n 또한 매우 커지는 경우를 상정한다면, 뉴턴-랩슨 방법은 사용하기 어렵다. 먼저 모수 $\boldsymbol{\theta}$ 의 차원 p 가 매우 큰 값을 갖는다고 하자. 예를 들어, $p = 10^6$ 인 경우를 고려하자. 이 경우 해시안 행렬의 크기는 $p \times p$ 이고, 뉴턴-랩슨 방법의 반복 과정에서 매번 이 행렬을 구하고 그 역행렬을 구하여 뉴턴-랩슨 방법을 적용하는 것은 비현실적이다. 즉, p 가 매우 커지는 경우라면, 뉴턴-랩슨 방법은 적용할 수 없는 방법이 된다.

이런 경우, 해시안 행렬을 사용하지 않는 경사하강(gradient descent) 방법을 대안으로 고려하게 된다. 경사하강법에서는 해시안 행렬을 적당한 상수로 대체한다. 즉, 다음과 같은 과정을 반복하여, 새로운 근사해 $\boldsymbol{\theta}^{(r+1)}$ 가 최적해에 근접하도록 한다.

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \lambda\left[\dot{L}\left(\boldsymbol{\theta}^{(r)}\right)\right], \quad r = 1, 2, 3, \ldots$$

이때 사용되는 상수 λ 를 ($\lambda > 0$) 학습률(learning rate)이라 한다. 경사하강법에서는 손실함수에 대한 해시안 행렬은 구할 필요가 없고, 그래디언트 벡터만 이용하여 계산한다. 경사하강법에서의 반복 횟수가 증가하면서 근사해가 최적해에 가까워지면 그에 따라 적절한 속도로 학습률을 줄여주어야 한다. 이를 위해 다양한 알고리즘들이 제안되었다. 경사하강법은 반복 과정에서의 계산 부담은 크게 줄일 수 있으나, 그 수렴 속도는 매우 느리고, 상당 횟수의 반복이 필요하다.

또 자료의 개수 n 이 매우 커지는 경우라면, 전체 자료 D 를 한꺼번에 이용하여, $L(\boldsymbol{\theta}) = L(D, \boldsymbol{\theta})$ 를 구하여 사용하는 것은 비현실적이 된다. 이런 경우, 전체 자료 D 를 한꺼번에 사용하는 대신, 전체 자료 D 를 적당한 크기로 분할한 여러 개의 배치자료 $\tilde{D}_r$, $r = 1, 2, \ldots, b$ 로 나누어 경사하강법을 적용한다. 이런 방법을 확률적 경사하강법(SGD)이라 한다. 다음의 (ii)와 (iii)을 1회 수행하는 것을 1 에폭(epoch)이라 하고, 확률적 경사하강법은 에폭을 반복하는 방법으로 수행한다.

(i) $\boldsymbol{\theta}^{(1)}$ 에 난수를 할당하여 초기화한다.

(ii) 전체 자료 D 를 랜덤 분할하여, $\tilde{D}_r$, $r = 1, 2, \ldots, b$ 생성 (여기서, $D = \bigcup_r \tilde{D}_r$)

(iii) $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \lambda[\dot{L}_r(\boldsymbol{\theta}^{(r)})]$, $r = 1, 2, \ldots, b$

(iv) $\boldsymbol{\theta}^{(1)}$ 에 $\boldsymbol{\theta}^{(b+1)}$ 을 할당한다.

(v) 위 (ii), (iii), (iv)를 반복한다.

이때, $L_r(\boldsymbol{\theta}) = L(\tilde{D}_r, \boldsymbol{\theta})$ 이고 $\dot{L}_r(\boldsymbol{\theta}) = \dot{L}(\tilde{D}_r, \boldsymbol{\theta})$ 이다. 배치자료의 크기 $|\tilde{D}_r|$ 를 배치의 크기(batch size)라 한다. 자료의 개수와 배치의 크기 s 사이에는 $n \approx b \cdot s$ 인 관계가 있다.

통계학에서 전통적으로 사용하여 온 최적화 방법은 뉴튼-랩슨 계열의 최적화 방법들이다. 뉴턴-랩슨 계열의 최적화 방법들을 매우 큰 자료와 모수의 개수가 매우 큰 모형들에 대하여 적용하기에는 현실적인 어려움이 있다. 때문에 심층신경망에서는 뉴턴-랩슨 방법 대신 확률적 경사하강법을 사용한다. 구체적인 함수 최적화

알고리즘에 차이가 있을 수 있으나, 통계적 모형이나 심층신경망 모두 손실함수를 최소화하는 방법으로 적당한 모수 $\boldsymbol{\theta}$를 찾는 과정을 수행한다는 공통점을 갖는다.

심층신경망이 기존의 통계적 방법과 구별되는 특징은 만능변환을 이용한다는 점이다. 깊이가 3인 다층퍼셉트론의 경우 $\boldsymbol{p} = [[[\boldsymbol{x}A]B]C]_*$ 와 같이 만능변환의 형태로 모수 $\boldsymbol{p}$를 모형화 한다. 이때 사용되는 가중치 행렬 $\boldsymbol{\theta} = (A, B, C)$ 가 추정 모수이다. 다층퍼셉트론에 경사하강법을 적용하기 위하여, 그래디언트 벡터 $\dot{L}(\boldsymbol{\theta})$를 구하는 과정을 살펴보자. 먼저, $\boldsymbol{p}$를 구하는 과정을 $\boldsymbol{g} = [\boldsymbol{x}\, A]$, $\boldsymbol{h} = [\boldsymbol{g}\, B]$, $\boldsymbol{p} = [\boldsymbol{h}\, C]$ 라고 단계별로 나누어 표현하자. 이에 다음과 같이 미분 연쇄법칙(chain rule)을 이용하여 그래디언트 벡터를 구한다.

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial C} &= \frac{\partial \boldsymbol{p}}{\partial C} \cdot \left(\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{p}} \right) \\
\frac{\partial L(\boldsymbol{\theta})}{\partial B} &= \frac{\partial \boldsymbol{h}}{\partial B} \cdot \left(\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{h}} \cdot \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{p}} \right) \\
\frac{\partial L(\boldsymbol{\theta})}{\partial A} &= \frac{\partial \boldsymbol{g}}{\partial A} \cdot \left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{g}} \cdot \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{h}} \cdot \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{p}} \right)
\end{aligned}$$

그래디언트 벡터를 구하는 과정에서, 위의 괄호 안의 과정과 같이, 손실함수 $L(\boldsymbol{\theta})$ 를, $\boldsymbol{p}$로 미분하고, 이를 다시 $\boldsymbol{h}$ 로 미분하고, 다시 $\boldsymbol{g}$ 로 미분하는 형태로, 만능변환의 마지막 단계의 변수에서 시작하여 앞 단계의 변수로 순서를 거슬러 가며 미분하는 과정을 반복한다. 이를 역전파(back propagation) 알고리즘이라 한다. 역전파 알고리즘은 만능변환의 특징을 이용하여 미분의 연쇄법칙(chain rule)을 반복하여 적용하는 과정을 의미한다.

3.3. 다그룹분류에 대한 고려

이론적 관점에서 보면 이진분류가 가능하면 이진분류를 반복하는 방법으로 다그룹분류는 당연히 가능하다고 생각되므로, 다그룹분류의 중요성이 낮게 평가될 수 있으나, 실제 자료의 해석에서는 다그룹분류가 자주 등장하고 그 중요성 또한 높다.

이진분류는 이항분포에 대하여 일반화선형모형을 적용하여 로지스틱 회귀로 연결되는 통계학의 주요 논제 중의 하나이다. 이에 반하여 통계학의 주요 논제에서 다그룹분류는 자주 언급되지 않는다. R의 기본함수인 glm 함수는 다그룹분류의 문제를 처리하지 못한다. R에서 glmnet 이라는 매우 강력한 패키지가 있기는 하지만, 아쉽게도 Windows 용은 없다. 이런 이유로 다수의 R 사용자들이 다그룹분류 문제를 처리할 때는 보통 nnet 패키의 multinom 함수를 사용하거나, mnlogit 패키지를 사용하게 된다 (Hasan 등, 2016). 이런 이유로, softmax 함수를 이용하는 다그룹분류는 전통적인 통계학의 논제가 아닌 것으로 생각되는 경향이 있다. 이는 다항분포가 지수분포족에 속하기는 하지만, 자연지수분포족(natural exponential family)에 속하지 않기 때문에 (Morris, 1983), 일반화선형모형에서 일차적으로 고려하는 분포가 아니라는 점과 연결된다고 보인다.

그러나 softmax 함수는 일반화선형모형에서 다루는 이론의 틀에서 다음과 같이 간단히 도출된다. 표현의 단순화를 위해, 다항분포 중 삼항분포를 예로 들어, 포아송분포와 삼항분포의 관계로부터 softmax 함수가 도출되는 과정을 살펴보자. 세 개의 독립인 포아송 확률변수 y_k, $k = 1, 2, 3$ 이 있고, 각각의 평균은 λ_k 라고 하자. 만약 세 변수의 합이 u 라는 조건이 주어진 경우라면, 즉 $y_1 + y_2 + y_3 = u$ 라면, $\boldsymbol{y} = (y_1, y_2, y_3)$ 는, $\boldsymbol{p} = (p_1, p_2, p_3)$ 일때, 다항분포 $\mathrm{Mnom}(u, \boldsymbol{p})$ 를 따르고, 다음 관계가 성립한다.

$$p_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \lambda_3}, \quad k = 1, 2, 3$$

또 포아송분포의 연결함수는 로그함수이므로 λ_k 는 $\log(\lambda_k) = \eta_k$ 로 모형화된다. 즉, $\lambda_k = \exp(\eta_k)$ 이다. 이를 p_k 에 대입하면 softmax 함수가 얻어진다.

3.4. 가변수 변환 방법의 차이

회귀분석에서 명목변수를 설명변수로 사용할 때, 여러 명목을 나타내는 변수를 0과 1로 이루어진 여러 개의 가변수(dummy variable)로 변환하여 대입한다. 보통 가변수의 개수는 명목변수에 포함된 명목의 개수보다 하나 작게 잡는 방법이 사용된다. 여러 명목 중 하나를 제외한 각각의 특정 명목에서만 1이 되고, 다른 모든 명목에서는 0이 되도록 가변수를 설정하는 방법이 주로 이용된다. 대비(contrast)는 특정 명목들 사이의 차이에 관심이 있는 경우에, 관심 명목 사이의 차이를 효과적으로 표현할 수 있는 가변수 설정 방법이다.

신경망에서는 명목변수를 라벨(label)이라 부르고, 원핫인코딩(one-hot encoding) 방법으로 가변수를 생성한다. 원핫인코딩은 각 변수들을 특정 명목에서만 1이 되고 나머지 모든 명목에서 0이 되도록 잡는 방법이다. 원핫인코딩으로 생성된 변수의 개수는 명목의 개수와 동일하다. 통계학에서 상수항(y-절편)이 없는 회귀모형에서 가변수를 설정할 때는 원핫인코딩과 동일한 방법을 사용한다.

원핫인코딩과 통계학의 가변수 설정법은 사실은 동일한 방법이다. 통계학에서 명목변수를 설명변수로 사용하는 회귀분석에서 상수항이 있는 모형에서는 가변수의 개수를 명목의 개수보다 하나 작게 잡는 것은, 상수항과 가변수 사이의 선형독립성을 보존하여, 모수 추정량을 특정하기 위한 목적이다. 신경망에서와 같이 모수추정을 포기하고 정확한 예측값을 구하는 것만을 목적으로 한다면, 굳이 가변수의 개수를 명목의 개수보다 하나 줄일 필요 없고, 원핫인코딩 방법을 사용하면 된다. 신경망에서 명목변수를 원핫인코딩으로 변환한다는 것은, 주요 관심사항이 예측값 $\hat{\boldsymbol{y}}$을 구하는 데 있고, 추청된 가중치 $\hat{\boldsymbol{\beta}}$에 의미를 부여하여 사용할 것을 포기했다는 말과 같다.

(다변량)회귀모형 혹은 일반화선형모형의 종속변수 혹은 신경망의 출력변수로 명목변수가 사용된 경우에도 원핫인코딩을 사용하는 것이 모형의 단순한 표현에 도움이 된다. 예를 들어, 세 가지 명목으로의 분류문제의 경우라면, 원핫인코딩 방법으로 얻어진 세 개의 가변수 y_1, y_2, y_3를 종속변수로 사용하게 된다. 이때 $y_1 + y_2 + y_3 = 1$ 인 제약조건이 부과되므로, 세 개의 가변수 중 두 개만을 종속변수로 사용하여도 무방하고, 세 개의 가변수를 사용하는 것에 비하여, 불필요한 계산을 피할 수 있다는 이점은 있을 수 있다. 그러나 그 계산상의 이익은 상대적으로 경미하고, 오히려 모형과 계산 절차의 단순화라는 관점에서 불이익이 더 크게 될 것이므로 단순하게 원핫인코딩을 사용하는 것이 더 유리하다.

4. 결론

앞서 심층신경망의 대표적인 모형인 다층퍼셉트론과 일반화선형모형으로 대표되는 전통적인 통계모형화 기법의 주요한 차이는 만능변환의 이용 여부임을 살펴보았다. 만능변환이 제공하는 폭넓은 함수근사 능력이 다층퍼셉트론과 심층신경망의 성공에 중요한 역할을 한 것은 틀림이 없다. 그러나 심층신경망을 성공으로 이끈 보다 근본적인 특징으로, 블랙박스 성질을 기본 틀로 설정하여 얻게되는 자유로운 설명변수 대입 기능과, 대량의 설명변수와 수많은 사례(instance)를 포함한 대규모 데이터에 적용할 수 있도록 개발된 확률적 경사하강법을 들 수 있다.

전통적인 통계적 모형과 심층신경망의 차이는, 구성해야 할 모형의 차이에서 비롯된 원인 보다는, 해결해야 할 현실 문제의 차이에서 비롯된 원인이 더 크다고 보인다. 신경망은 대규모 데이터에서 변수선택의 폭을 넓게하여, 정확도 높은 예측값을 얻는 방향을 추구했고, 이 방향의 차이가 신경망과 통계학의 차이를 낳은 근본적인 원인이었다.

블랙박스 성질을 가진 심층신경망의 대두는, 통계학 학문의 전체 구도에서 더 큰 시사점을 준다고 할 것이다. 통계적 모형이나, 심층신경망 모두 미래에 발생할 수 있는 새로운 관측값에 대한 예측을 주요 관심대상으로 하고 있다. 통계학이 취한 접근법은, 관측된 자료를 이용하여 모집단의 특성을 잘 규정하면, 향후 새로이 관측될 미지의 값에 대한 예측도 잘 할 수 있을 것으로 보고 일차적으로 모집단의 특성을(즉, 모수를) 잘 특정하는데 주력하는 방법이었다. 이에 반하여, 심층신경망이 택한 접근법은 굳이 모집단의 특성을 특정

하여 예측값이 얻어지는 과정을 명확하게 밝히지 않더라도, 좋은 예측값을 얻을 수 있다는 점에 착안하여, 모집단의 특정과 그에 따른 해석을 포기하고, 더 우수한 예측값을 얻는 데 주력했다는 점이다. 이는 모집단의 특정을 우선시 해왔던 통계학의 접근법 전반에 대한 폭넓은 고찰을 요구하고 있다.

요인분석(factor analysis), 시계열분석법, 커널평활법(kernel smoothing method)과 같은 다양한 분석방법이 통계학의 영역에서 개발되고 발전되었다. 이런 전통적인 통계적 방법이 심층신경망과 결합되어 어떻게 발전하게 되었는지에 대한 깊이 있는 고찰이 추가로 필요하리라 보인다. 또, 확률적 경사하강법을 통계학의 입장에서 해석하는 연구가 추가로 필요하다고 생각된다. 여기에 최근 관심을 받고 있는 RNN 방법들과, 트랜스포머 등 언어모형에 구현된 통계 방법론 등을 검토하여 (Kim 등, 2024a; Kim 등 2024b), 통계학의 지평을 넓히고 통합적 틀을 구성하는 작업이 필요할 것이다. 더하여, 의료, 보건, 금융, 마케팅 등 다양한 인접 학문 영역에로의 활용 연구가 이어지길 기대한다 (Hwang 등, 2024).

References

Breiman L, Friedman J, Stone CJ, and Olshen RA (1984). *Classification and Regression Trees* (1st ed), Chapman & Hall, New-York.

Cybenko G (1989). Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems*, **2**, 303–314.

Hastie T, Tibshirani R, and Friedman J (2001). *The Elements of Statistical Learning*, Springer-Verlag, New-York.

Hasan A, Wang Z, and Mahani AS (2016). Fast estimation of multinomial logit models: R package mnlogit, *Journal of Statistical Software*, **75**, 1–19.

Hwang IJ, Kim HJ, Kim YJ, and Lee YD (2024). Generalized neural collaborative filtering, *The Korean Journal of Applied Statistics*, **37**, 311–322.

Kim HJ, Kim YJ, Jang K, and Lee YD (2024a). A statistical journey to DNN, the second trip: Architecture of RNN and image classification, *The Korean Journal of Applied Statistics*, **37**, 553–565.

Kim YJ, Hwang IJ, Jang K, and Lee YD (2024b). A statistical journey to DNN, the third trip: Language model and transformer, *The Korean Journal of Applied Statistics*, **37**, 567–582.

Morris CN (1983). Natural exponential families with quadratic variance functions: Statistical theory, *The Annals of Statistics*, **11**, 515–529.

Venables WN and Ripley BD (2002). *Modern Applied Statistics with S* (4th ed), Springer-Verlag, New-York.

Received July 31, 2024; Revised August 09, 2024; Accepted August 12, 2024

심층신경망으로 가는 통계 여행, 첫 번째 여행:

회귀모형에서 심층신경망으로

김희주[a], 황인준[a], 김유진[a], 이윤동[1,a]

[a]서강대학교 경영학부

요 약

최근 인공지능과 심층신경망에 대한 언급 없이 통계학을 이야기하기 어려운 시대가 되었다. 인공지능과 심층신경망의 발전은 통계학의 주요 연구 성과가 이루어 낸 결과이기도 하지만, 현대의 통계학과 인공지능은 사뭇 다른 방법인 것처럼 생각되기도 한다. 그 주요 원인은 통계학 교육과정이 시대에 맞게 변화하지 못한 데 따른 것으로 보인다. 본 논문에서는 통계학 교육의 확장과 발전의 틀을 마련하기 위하여, 심층신경망 그 중에서도 다층퍼셉트론과 회귀분석의 관계를 통계학의 관점에서 살펴보고, 그 공통점과 차이점을 살펴본다.

주요용어: 심층신경망, 분류, 다층퍼셉트론, 회귀분석, 일반화선형모형

[1]교신저자: (04107) 서울시 마포구 백범로 바오로관 804호, 서강대학교 경영학부. E-mail: widylee@sogang.ac.kr